SUPPLEMENTARY INFORMATION

Selective Insensitivity to Global Versus Local Linguistic Context in Speech Produced by Patients With Untreated Psychosis and Positive Thought Disorder

Sharpe *et al.*

Table of Contents

1. Data Collection: Additional Methodological Details 2

2. Background: Context and Lexical Probability 8

3. Estimating Lexical Probability in Natural Language: 20

4. The Present Study: Extraction of Lexical Probability Values Under Multiple Conditions . 29

5. Preparation of Data for Statistical analysis 38

6. Assumptions for Hypothesis Testing and Ruling Out Artifactual Explanations for the Group*ContextWindowSize Interaction 40

7. Complete Results Tables for Analyses Reported in the Main Manuscript 47

8. Additional Statistical Analyses 51

9. References 65

1. Data Collection: Additional Methodological Details

1.1 Assessment of general cognitive function: Additional details on tasks and scoring

As described in the main manuscript, general cognitive function was assessed in all participants, using three cognitive tasks as measures of (a) working memory and processing speed; (b) semantic memory and executive function, and (c) non-verbal executive function. Here, we provide additional details of how these assessments were carried out and scored.

(a) Working memory and processing speed

Task: A modified version of the Digit-Symbol Substitution Test (DSST; original: Wechsler, 1981; modified: Palaniyappan, Al-Radaideh, Mougin, Gowland, and Liddle, 2013).

Instructions: Participants were asked to match symbols to numbers based on a given key. Participants completed both written and oral portions of the task.

Scoring: Participants scores reflected the time taken to complete each portion of the task. We averaged each participant's written and oral scores to form a single composite score.

(b) Semantic memory and executive function

Task: Category Fluency Task (Neill, Gurvich, & Rossell, 2014, for review).

Instructions: In this task, participants were asked to name as many exemplars of a particular semantic category (here, for the category "animals") as possible.

Scoring: The number of (appropriate) exemplars that the participant produced was tallied.

(c) Non-verbal executive function

Task: Part B of the Trail-Making Test (TMT; Arbuthnott & Frank, 2000)

Instructions: Participants traced a line through consecutive numbers and letters, alternating between numbers and letters (1-A-2-B-3-C-and so on).

Scoring: The time taken to complete this connect-the-dots-style pattern correctly was recorded.

1.2 Elicitation of speech samples and details of transcription

To obtain the speech data for this study, participants described three pictures from the Thematic Apperception Test (Murray, 1943) for one minute each. Participants were given the following standardized instructions:

"I am going to show you some pictures, one at a time. When I put each picture in front of you, I want you to describe the picture to me, as fully as you can. Tell me what you see in the picture, describe what you see in this image, and what you think might be happening."

If, for any given picture, the participant stopped speaking before their minute was up, the experimenter prompted them to go on or give more detail (e.g. "What else do you see?").

Speech recordings were manually transcribed by trained research assistants, with

punctuation included per the transcriber's intuitions. Each transcript was checked by an additional trained researcher. This initial transcript was used for the purpose of clinical ratings of thought disorder, as described next (Supplement 1.3). Note that, *before* feeding the transcripts into GPT-3 to obtain lexical probability values, we carried out several extra steps to standardize the transcripts (see Supplement 4.1), including stripping any non-essential punctuation.

1.3 Assessment and operationalization of positive thought disorder ("Disorganization") and negative thought disorder ("Impoverishment")

Thought disorder ratings were conducted using the Thought and Language Index (TLI; Liddle, Ngan, Caissie, Anderson, Bates, Quested, White, & Weg, 2002). These ratings were based on the initial speech transcriptions and were completed by a single trained graduate-level research assistant under the supervision of a research psychiatrist. Both were blinded to patient status.

For the purpose of statistical analysis, we used the TLI assessment to compute separate summary measures of positive thought disorder ("Disorganization" score) and negative thought disorder ("Impoverishment" score) in each participant. To operationalize positive thought disorder ("Disorganization" score), we took the sum of ratings for the following TLI items: Looseness, Peculiar Word Use, Peculiar Sentence Construction, and Peculiar Logic. To operationalize negative thought disorder ("Impoverishment" score), we took the sum of ratings for the following TLI items: Poverty of Speech and Weakening of Goal.

1.4 Generation of an unrelated picture description

As described in the main manuscript, for our first statistical analysis (Results, Section 1), we sought a baseline measure that would reflect the probability of every word in each speech segment, given the same number of preceding words of context, but in the absence of any type of contextual structure (P(word | UnrelatedScrambledContext)). This allowed us to exclude the possibility that any group differences in lexical probability were driven by inherent differences in the particular words participants used.

For this purpose, we needed a speech sample for an unrelated picture from which to randomly select words to generate the scrambled context. We simulated this picture description using the ChatGPT (OpenAI, 2023) tool. However, we emphasize that the specific text used to generate the scrambled contexts is not critical; the goal was simply to disturb the structure and content of the prior context for each word.

Here, we describe the instructions we gave to ChatGPT to generate this unrelated picture description, and we also provide the simulated picture description itself. In Section 4.3 of this Supplement, we describe how we sampled random words from this unrelated picture description in order to compute P(word | UnrelatedScrambledContext).

We gave ChatGPT the following prompt to elicit the speech sample:

"I am carrying out a study in which I am asking participants to describe black-and-white

pictures.

The instructions to the participants are as follows:

"I want you to describe the picture to me, as fully as you can. Tell me what you see in the picture, describe what you see in this image, and what you think might be happening."

I would like you pretend that you are an average participant in this study, and you are asked to describe a picture of a boy who is sitting with his elbows on the table and in front of him is a violin and bow. He has his head in his hands and is staring sadly at the violin.

Please provide a transcript of a natural utterance that you might produce when asked to describe this picture. You may elaborate on the kinds of things that might be in this picture, but please constrain what you say to what someone might realistically say about a single black and white picture.

It should sound like you are actually talking - not writing.

The language should sound colloquial without a lot of fancy words - the kind of thing that a regular person might say.

The utterance should be 255 words long."[1]

Simulated speech sample provided by ChatGPT

[1] Note that this length was determined by taking the length, in words, of our longest speech sample.

“Okay, so there's this black-and-white photo, right? And there's this kid, maybe around 10 or 12 years old, sitting at a wooden table. He's got his elbows on the table and he's leaning forward a bit. He's got short, messy hair, and he's wearing a casual button-up shirt with the sleeves rolled up.

Now, right in front of him on the table is a violin and a bow. The violin looks nice and well taken care of. The bow is just lying there next to it, and you can see the strings on it. The black and white in the picture makes everything stand out.

The kid has his head in his hands, like he's really sad or something. He's staring at the violin with this downhearted look on his face. It's like he's been crying or is about to cry, you know? It's kinda touching to see.

In the background, there's a window with some light coming in. It makes some shadows in the room and adds to the whole sad vibe. You can't really see what's outside, but it might be a cloudy day or something.

So, yeah, it looks like the kid is having a tough time with the violin, or maybe something bad happened related to it. The whole thing feels pretty emotional, you know? Like, it really makes you feel for the kid and what he's going through.”

“Okay, so there's this black-and-white photo, right? And there's this kid, maybe around 10 or 12 years old, sitting at a wooden table. He's got his elbows on the table and he's leaning forward a bit. He's got short, messy hair, and he's wearing a casual button-up shirt with the sleeves rolled up.

Now, right in front of him on the table is a violin and a bow. The violin looks nice and well taken

care of. The bow is just lying there next to it, and you can see the strings on it. The black and white in the picture makes everything stand out.

The kid has his head in his hands, like he's really sad or something. He's staring at the violin with this downhearted look on his face. It's like he's been crying or is about to cry, you know? It's kinda touching to see.

In the background, there's a window with some light coming in. It makes some shadows in the room and adds to the whole sad vibe. You can't really see what's outside, but it might be a cloudy day or something.

So, yeah, it looks like the kid is having a tough time with the violin, or maybe something bad happened related to it. The whole thing feels pretty emotional, you know? Like, it really makes you feel for the kid and what he's going through."

2. Background: Context and Lexical Probability

2.1 Hierarchical encoding of linguistic and non-linguistic dependencies at different temporal scales constrains the lexical properties of individual upcoming words

Language is a sequential code used to communicate our thoughts and experiences. Accordingly, any given sequence of words within a context encodes multiple linguistic and non-linguistic dependencies that together influence the linguistic and non-linguistic properties of the next word in the sequence. These dependencies include our knowledge about relationships between single words, syntactic and semantic constraints, and our knowledge about real-world

topics and schemas.

To give some intuition, consider the following example:

"Sarah had spent the entire afternoon preparing for the big presentation, making sure every detail was perfect. But as she stood in front of the room, she suddenly realized she had forgotten her… notes"

Given the preceding context, the lexical properties of the word "notes" is a plausible continuation.

Some of the dependencies that influence this word are as follows:

(a) Syntactic constraints: The structure ("her __") constrains for a noun.

(b) Semantic selection restrictions: Semantically, the phrase "had forgotten her ___" constrains for an object that can be forgotten.

(c) Topic and schema-based knowledge: Most people are familiar with the act of giving presentations and know that, in the real world, presenters often use notes to stay organized and remember key points.

From this example, it should also be clear that different types of contextual dependencies and knowledge tend to operate over different time spans. Short-term syntactic dependencies and semantic selection restrictions operate within short phrases, while whole messages, schemas, and topics are often encoded across multiple sentences.

To characterize the timescales at which these different contextual constraints operate, researchers often appeal to a hierarchical structure. The notion of hierarchy is invoked in many different ways in language research (e.g. syntactic phrase structures). Here, however, we refer to a hierarchy of *levels of representation* which encodes different types of linguistic dependencies at increasingly long timescales. At lower levels of the hierarchy (the shortest timescales), lexical representations encode the linguistic forms (orthographic and phonological) of single words. At middle levels of the hierarchy, syntactic and semantic dependencies interact closely to specify the roles of "who does what to whom" (to form linguistic event structures); at higher levels of the hierarchy, discourse structures link to our real-world knowledge of whole schema and topics.

Because successively higher levels of this hierarchy encode knowledge that spans increasingly longer timescales, each individual content word, encoded at the lowest levels of the hierarchy, will be constrained by contextual dependencies encoded at all levels above it. Therefore, in the example above, the word, "notes" is constrained by semantic and syntactic knowledge about event structures, encoded at middle levels of the hierarchy (e.g. <woman forgot notes>), as well as by real-world schema and topic-relevant knowledge about {giving presentations}, encoded at higher levels of the hierarchy. In other words, many constraints encoded in the prior context "bottom out" to inform the lexical properties of each subsequent word. (Note that, compared to content words, function words are likely to be more constrained by semantic and syntactic dependencies encoded at mid-levels of the hierarchy, and less by real-world schema or topic knowledge, encoded at higher levels of the hierarchy.).

2.2 Why do contextual constraints influence *lexical probability*, and why does lexical probability play such a crucial role in the production of coherent language in healthy adults?

The fact that the properties of a single word (lexical item) are constrained by linguistic and non-linguistic dependencies encoded within the prior linguistic context (the preceding sequence of words) is not simply of academic interest. Being able to *exploit* these contextual dependencies is essential for effective real-time communication. During language processing, the full sequence of words, and by extension, the full hierarchy of representations, is not available all at once. Instead, language is both produced and comprehended incrementally, one word at a time. As a result, each word acts as a critical *bottleneck* during communication, constraining the flow of information from one person to another. This bottleneck faces significant pressure to optimize efficiency due to three key challenges: the rapid pace of language production and comprehension, the inherent ambiguity of the linguistic code, and the presence of noise in both the internal and external communicative environment.

Therefore, to ensure efficient communication and minimize errors, both the producer and the comprehender must rapidly draw upon their shared linguistic and non-linguistic knowledge, encoded within the prior context, to facilitate lexical processing; that is, all these sources of knowledge must *interact* with one another during incremental real-time language processing.

For the producer, the most effective way to convey their intended message is to use the prior context to make each word that they produce as predictable as possible, given the constraints of the preceding context, thereby making it easier for the comprehender to process. Conversely, for the comprehender, the most effective strategy for quickly and accurately

recognizing each incoming word is to leverage the constraints of the prior context to predict ahead. It is these rational principles of communication that explain why lexical probability plays such a central role in language processing Levy, 2008.

Lexical probability simply describes the predictability of a word in relation to its preceding context (the sequence of words that precede it) as a function of all the linguistic and non-linguistic knowledge encoded in that context. It is therefore an inherently *holistic* measure that captures *multiple* types of dependencies, which *interact* closely. It is easier to process words that are more predictable than less predictable. Therefore, to produce speech that is optimally coherent and easily understandable, a rational *producer* should continuously track the sequence of words they have already produced (i.e., the context available to the comprehender) and adjust the probability of each subsequent word accordingly. Similarly, a rational *comprehender* should use whatever context is available to them to incrementally build hierarchical representations of the unfolding input and use this information to estimate the probability of the next word they may encounter so that its processing is facilitated to the degree to which it has been predicted. Of course, the comprehender will never know precisely what word will come next and so their predictions will be probabilistic. However, so long as multiple potential lexical representations are pre-activated in parallel, each with a strength that reflects its probability given the context, then, on average, the benefits of predictive facilitation should outweigh any costs (see Kuperberg & Jaeger, 2016, for discussion).

Importantly, these intuitions about the central role of lexical probability in communication are supported by a large body of research. One of the most consistent findings in psycholinguistic and neurolinguistic research over the past few decades is that, in healthy adults,

the probability of each incoming word is an extremely robust predictor of both behavioral response times and neural activity during language processing (see Kuperberg & Jaeger, 2016); that is, the more probable (predictable) a word is given the context that precedes it, the faster it is to process and the less neural activity it evokes. This effect of lexical probability on lexical processing is observed regardless of the entropy/constraint of the prior context (see Section 3.1 below). Moreover, the effect is robust regardless of whether lexical probability is estimated based on human Cloze norms (for behavioral evidence, see e.g. Rayner & Well, 1996; Brothers & Kuperberg, 2021; for neural evidence, see e.g. Kutas & Hillyard, 1984; DeLong, Urbach & Kutas, 2005; Brothers, Morgan, Yacovone, & Kuperberg, 2023), or based on the outputs of large language models (for behavioral evidence, see e.g. Shain, Meister, Pimentel, Cotterell, & Levy, 2024; Boyce & Levy, 2023; for neural evidence, see e.g. Michaelov et al., 2024; Szewczyk & Federmeier, 2022).[2]

[2] Several studies examining the relationship between lexical probability and behavioral and neural response measures apply a log-transformation (specifically, the negative logarithm) to lexical probability values. The negative logarithm of a probability value is referred to as "surprisal", with more probable words being less surprising. For some researchers, this logarithmic transformation is primarily a methodological choice to meet the assumptions of general linear models, as it effectively "spreads out" data points near zero (see Section 6.1(b) below). This consideration is particularly relevant for studies that derive lexical probability estimates from LLMs, which, unlike Cloze measures, yield meaningful estimates for very low-probability data points (see Section 3.3 below). In the psycholinguistic comprehension literature, however, other researchers place greater theoretical emphasis on this transformation, highlighting surprisal as an information-theoretic measure (Shannon & Weaver, 1949; e.g., Hale, 2001; Levy, 2008; Smith & Levy, 2013). There is ongoing debate about the precise linking function between lexical probability and behavioral and neural measures of word-by-word processing. Some have argued that the negative logarithm of the lexical probability, i.e. surprisal, is always a best predictor of response times (Shain, Meister, Pimentel, Cotterell, & Levy, 2024). However, others have shown that for probability values between 0.1-1, there is a linear relationship between raw lexical probability values and both response times (Brothers & Kuperberg, 2021) and evoked neural activity (Szewczyk & Federmeier, 2022).

2.3 The use of ContextWindowSize as a continuous measure that *proxies* the influence of on the lexical probability of content words

To sum up, communication is a dynamic and cooperative process in which both the producer and the comprehender continually draw on their linguistic knowledge about dependencies encoded at multiple timescales within the preceding context. Producers can use this knowledge to influence the probability of each word they produce.

In the present study, we aimed to compare the timescales over which patients and controls used context to influence the lexical probability of content words during natural speech production. To do this, we used GPT-3 to estimate the probability of each content word produced based on varying context lengths ("context window sizes"), ranging from very local (a single preceding word) to very global (up to 50 prior words). (Note that to extract these estimates, we provided GPT-3 with the full contexts — both content words and function words, see Supplement, Section 5.1 below).

Broadly speaking, context window size (the number of prior words) can be considered a proxy for the *depth* of the representational hierarchy that the brain uses to influence lexical probability. Estimates of lexical probability that are based only on a *local context* will reflect the use of information encoded at mid-levels but not higher levels of the representational hierarchy. Conversely, estimates of lexical probabilities that are based on the full *global context* will generally reflect the use of information encoded at *both* mid-levels *and* higher levels of the

hierarchy.[3] To give some intuition, let us return to the speech sample above (repeated here for convenience):

"Sarah had spent the entire afternoon preparing for the big presentation, making sure every detail was perfect. But as she stood in front of the room, she suddenly realized she had forgotten her… notes"

If we were to estimate the probability of the content word, "notes", based solely on the preceding four words, "she had forgotten her…" (the local context), this estimate would depend primarily on syntactic and semantic dependencies encoded at mid-levels of the hierarchy. Note that these semantic and syntactic dependencies *interact* closely with one another to influence the probability the next word. Based on these dependencies, the probability of "notes" will be greater than the adverb, "quickly," which is less expected, given the local syntactic structure. It will also be greater than that of a word like "clouds," which violates the semantic restrictions of the preceding verb, "forgotten." However, based solely on this local context, the probability of "notes" would be no greater than that of any other noun that can be forgotten, such as "pajamas" or "money."

If, however, we were to estimate the probability of "notes" based on the entire preceding

[3] Throughout the manuscript, we use the term "global context" to refer to the use of a large context window size (a large number of words in the prior context) to estimate lexical probability, and "local context" to refer to the use of a smaller context window size to estimate lexical probability. When defined in this way, "global context" *subsumes* "local context." To avoid confusion, we use the term "global sources of information" to refer more specifically to the additional types of information offered by global contexts (e.g., schema-relevant information, topics, discourse structure, and relevant visual cues), which are thought to be encoded at higher levels of the representational hierarchy.

global context, this estimate would depend not only on syntactic and semantic dependencies encoded at mid-levels of the hierarchy (as described above), but *also* on schema- and topic-relevant real-world knowledge encoded at higher levels of the hierarchy. Once again, *all* these dependencies must interact closely with one other to influence the probability of the upcoming word. Because "notes" aligns with schema- and topic-relevant information at these higher levels, its probability will be increased (upweighted) relative to its probability based solely on the local context. In contrast, the probabilities of schema-inconsistent words like "pajamas" or "money" will be decreased (downweighted) relative to their probabilities based only on the local context.[4]

On the other hand, it is important to recognize that the relationship between context window size and depth of the representational hierarchy is a broad approximation only: There is *no simple one-to-one correspondence between the size of the context window and the specific types of linguistic and non-linguistic dependencies that influence lexical probability*. A window size of up to 5–7 words will often capture the semantic and syntactic information necessary to specify "who does what to whom" — simple event structures encoded within clauses or short sentences (e.g. Mollica et al, 2020), but these types of dependences can operate over longer distances. Conversely, discourse relationships that establish entire topics and schemas will often

[4] Note that these differences between estimates of lexical probability based on global vs. local context are primarily expected for *content words*. The probability of *function words* will also be influenced by local contexts. However, as noted above in Section 2.1, function words are less constrained by global sources of information, such as broader topic and schema-relevant knowledge. Therefore, in healthy adults, the difference between estimated probabilities of function words given local context P(function word | local context) and global context P(function word | global context) is likely to be less pronounced than the corresponding difference for content words. It is for this reason that, in the present study, we focused our statistical analyses on the probabilities of *content words* to test the main hypothesis: that patients' sensitivity to global (vs. local) context is disproportionately reduced.

operate over longer distances. However, they can also be established by a single informative word or phrase. Therefore, there is *no fixed inflection point of context window size* at which one can say that producers transition from relying solely on middle levels to using both mid- and higher-level representations of the hierarchy to inform lexical probability; boundaries cannot be easily mapped onto predefined context window sizes. Instead, they vary widely across typical speech samples, reflecting the inherent flexibility in how events and topics are organized.

For this reason, in our analyses, we treated the local-global distinction as a graded *continuum* rather than as a binary division. In other words, we assume that, averaged across all speech samples, as the context window size increases, more information relevant to a particular intended message (and particular associated schema/topic) is available, and so each successive content word will become, on average, increasingly more predictable. This type of *graded* increase in lexical probability with increasing context window size has been observed in previous work (e.g. Giulianelli & Fernandez, 2021; Sharpe, 2023). It was clearly evident in the present study: As shown in Figure 3, in both healthy adults and people with schizophrenia, as context window size increased, the lexical probability of content words steadily increased (although at a slower rate in patients than controls, as we discuss next).

2.4 Why should a relative insensitivity in the use of global (vs. local) context to modulate lexical probability serve as an objective index of positive thought disorder in schizophrenia?

From the discussion above, it should be clear that the production of coherent speech critically depends on the producer's ability to use all available context to modulate the

probability of each word they produce.[5] It therefore follows that if a producer fails to account for the prior context, their speech will become more disorganized and incoherent. At the extreme, a failure to account for *any* context would render speech completely incoherent and unintelligible. This, however, is typically not seen in people with positive thought disorder. Instead, as noted in the Introduction of the main manuscript, clinical phenomena such as tangentiality and derailment seem to suggest a relative preservation in patients' ability to use local context, but a reduced sensitivity to more global context.

To test this hypothesis of this we tested for *interaction* between Group (FEPs vs. HCs) and ContextWindowSize. ContextWindowSize was treated as a *continuous* variable because, as explained above (Supplement, 2.3), there is no one-to-one relationship between context window size and depth of hierarchy/type of linguistic dependencies, with boundaries varying widely across speech samples.

This analysis confirmed a Group*ContextWindowSize interaction. This interaction was graded; that is, the difference in lexical probability between FEPs vs. HCs increased gradually as the ContextWindowSize increased. This was the case both in linear and logarithmic space. In other words, there was no specific context window size at which the difference between patients and controls first became specific window size in which the emerged. Again, this is not surprising, given that thought disorder encompasses diverse phenomena, as captured by the major clinical rating scales, and so there is likely to be even more heterogeneity in the precise

[5] Indeed, the very success of large language models, trained to predict missing words based on large context window, as discussed below in Section 3.3, speaks to the importance of lexical probability in producing coherent language.

relationship between representational hierarchy and ContextWindowSize in the patient group.

These findings confirm that patients with positive thought disorder exhibit a *disproportionate* insensitivity to global context, relative to their ability to use more local context. Because context window size serves as a broad approximation for the types/levels of representation that influence lexical probability, these findings these findings can be interpreted as evidence of a *selective* impairment in maintaining and/or using representations at *higher vs. lower levels* of the representation hierarchy. That is, lexical probabilities in patients' speech benefitted less from the weighting typically conferred by these higher-level representations.

On the other hand, it is important to recognize that there is no straightforward one-to-one mapping between context window size and the precise linguistic or non-linguistic representations used to inform lexical probability (see Supplement 2.3). Therefore, the current data do not allow us to draw definitive conclusions about the specific types of dependencies (i.e., levels of the linguistic hierarchy) that patients can or cannot use during language production (see Supplement 2.4). To better understand the nature of these representations, we must turn to models that are more transparent than GPT (and other current LLMs).

While the strength of LLMs lies in their ability to output lexical probability estimates that closely correlate with human estimates (e.g., Sharpe, 2023; Hao, Mendelsohn, Sterneck, Martinez, & Frank, 2020; Szewczyk & Federmeier, 2022) and predict behavioral (e.g. Shain, Meister, Pimentel, Cotterell, & Levy, 2024; Boyce & Levy, 2023) and neural (e.g., Michaelov et al., 2024; Szewczyk & Federmeier, 2022) measures of lexical processing (see Supplement, 3.5), their "black box" nature limits our understanding of the representations underlying these probabilities; that is, their parameters are not inherently interpretable.

As we explain in the Discussion of the main manuscript, in order to fully characterize high- and lower-level atypicalities in schizophrenia, and explore individual differences in patients' sensitivity to different types of global context that can be linked to specific subtypes of positive (and negative) thought disorder, it will be necessary to develop transparent and interpretable models capable of directly probing these high-level representations. This will also require larger samples and greater variance in positive (and negative) thought disorder symptoms.

3. Estimating Lexical Probability in Natural Language: From Cloze Probability to Large Language Models

3.1 Estimating lexical probability: General principles and relationship with contextual entropy and constraint

Obviously, we can never know the precise probabilities represented in speakers' or listeners' minds at any specific moment during language production or comprehension. However, it is possible to *estimate* the probability of a given word, in any language sample, based on any amount or type of context that precedes the word. Several different methods exist for estimating these lexical probability values. However, they all share the same fundamental assumption: Data collected from many individuals can provide an accurate estimate of the lexical probability of a word produced by a single individual. This phenomenon is referred to as the "wisdom of the crowd" effect (Galton, 1907; Stroop, 1932).

Thus, if we provide multiple humans, or a Large Language Model (trained on many

humans' language output), with the context introduced above—"*Sarah had spent the entire afternoon preparing for the big presentation, making sure every detail was perfect. But as she stood in front of the room, she suddenly realized she had forgotten her …*"—then we will obtain a distribution of probabilities over multiple possible words (e.g. 50%: "notes", 30%:"slides", 10%:"handouts", etc.).[6] To determine the probability of the word that is actually produced, we simply look its probability value up from this probability distribution. For example, if the produced word is "notes," the lexical probability of "notes" given this context would be 0.5.

We note that, in addition to deriving the probability of the word that is actually produced/observed (lexical probability), it is also possible to derive other measures that characterize the *context, regardless* of the word that is actually produced/observed. For example, we can compute the *entropy* of this probability distribution, which characterizes our uncertainty about upcoming continuations based on the context, before observing any given word: High entropy (a wide, flat distribution) indicates greater uncertainty about the next word, and low entropy (a narrow, tall distribution) indicates less uncertainty. The entropy of the prior context is sometimes approximated by a measure known as *contextual constraint,* which is simply defined as the probability of the *most* probable completion following a given context, again regardless of the word that is actually produced/observed. For example, in the example above, the most probable completion is "notes" (50%) and so the constraint of this context is 0.5, even if the next word produced was "handouts".

[6] Recall that this follows from the theoretical assumptions discussed above That is, to maximize their chance of recognizing each new incoming word, a rational comprehender should generate *parallel probabilistic predictions* about this upcoming word, with each expected lexical representation pre-activated with a strength that reflects its probability.

Although contextual entropy and contextual constraint may also play a role in communication, there is strong evidence that the probability of the particular word that is actually produced by far the most robust predictor of behavioral and neural indices of word-by-word processing, and that it predicts these measures *regardless* of the constraint of the prior context, or the probability of any alternative competing predictions. For example, if a context constrains strongly for a specific word that is then disconfirmed by an unexpected bottom-up input (e.g. *The young nervous paratrooper jumped out of the...*; expected: *plane* vs. *chair*), both response times and evoked neural responses are no larger than when the same unexpected input appears in a low-constraint contexts where no individual lexical item was strongly pre-activated (e.g. *The tired movie maker was sleeping in the chair*) (behavioral evidence: Frisson et al., 2017; neural evidence: Kuperberg, Brothers, & Wlotko, 2020). Similarly, response times and evoked neural activity to *partially predictable* words are not affected by whether the context pre-activates a higher probability unrelated word. For example, in the sentence "*Stephen wanted to do something special for his girlfriend. He decided to make her a hand-made card. On it, he drew some...flowers*" (with the higher-probability unrelated word being *hearts*), both response times and evoked responses to *flowers* are comparable to those in a low-constraint context like "*Stephen always doodled in class. He took out a fresh sheet of paper. On it, he drew some…flowers*" (behavioral evidence: Luke & Christianson, 2016; Frisson, Harvey, & Staub, 2017; neural evidence: Brothers, Morgan, Yacovone, & Kuperberg, 2023).

It is for this reason that, similar to most psycholinguistics research, we chose to use the lexical probability of the produced "target" word as our measure of interest.

3.2 Human crowd-sourced estimates: Cloze probability

Traditionally, and in most psycholinguistic studies (at least until recently), the probabilities of lexical continuations have been estimated using a method known as the Cloze procedure (Taylor, 1953). In this procedure, a large number of participants are presented with a context that omits the "critical word" of interest and are asked to "fill in the blank" with the word they think is most likely to come next. The participants' responses are then aggregated to construct a probability distribution. For example, 100 participants may be given the context, "*Sarah had spent the entire afternoon preparing for the big presentation, making sure every detail was perfect. But as she stood in front of the room, she suddenly realized she had forgotten her ...*". If 50 of the 100 participants complete the sentence with the word ("notes"), then the lexical (Cloze) probability of "notes" is estimated to be 0.5.

Because Cloze captures probabilistic data directly from humans, it has been considered the "gold standard" in psycholinguistic research for quite some time. However, collecting Cloze completions is quite time-consuming and it is therefore less appropriate for estimating the probability of every single word in large, complex datasets (although see Luke & Christianson, 2018). This makes it particularly problematic for estimating precise probabilities of very low probability words: Millions of samples per context would be needed to obtain reliable estimates.

3.3 Corpus-based NLP methods and Large Language Models (LLMs)

An alternative to human-generated Cloze values is to leverage large corpora in combination with Natural Language Processing (NLP) techniques to estimate lexical

probabilities. While human-generated values depend on explicit crowd-sourced continuations, these methods estimate lexical probability based on implicit statistical patterns embedded within large amounts of naturally occurring, human-produced language data.

Until recently, the main NLP methods used to estimate lexical probability was frequency-based techniques, such as n-gram models. N-gram models estimate the conditional probability of a word given a context of the previous n-1 words. These probabilities are computed based on the frequency of each n-gram in a training corpus. However, a significant limitation of this method is that the number of possible n-grams increases exponentially as n becomes larger, which means that there will be many n-grams that never occur in the training corpus. While smoothing techniques exist to mitigate zero probabilities for these unseen n-grams (see, e.g., Chen & Rosenfeld, 2000, for review), estimated probabilities for n-grams larger than bigrams (n=2) or trigrams (n=3) are unlikely to be reliable. This is because, as discussed above in Section 2.1, such short contexts typically capture only a small portion of the information that can influence the probability of an upcoming word, i.e. the influence of longer-range global context is ignored.

More recently, however, the development of large language models (LLMs) has provided a powerful new way of estimating lexical probability on any given word, based on its *full* available context. These deep learning models employ neural network architectures, trained on vast amounts of text data, to "learn" statistical regularities in language, including both local patterns as well as long-range global dependencies. Most famous among these are the transformer models. The transformer is a non-recurrent deep learning architecture that can capture relationships between "tokens" (i.e. words, subwords or characters) and long, extensive contexts. Transformer-based models include the GPT (Generative Pre-trained Transformer) family of

models developed by OpenAI (Radford, Wu, Child, Luan, Amodei & Sutskever, 2019; Brown et al. 2020),[7] BERT (Bidirectional Encoder Representations from Transformers; Devlin, Chang, Lee & Toutanova, 2018), and LLaMA (Large Language Model Meta AI; Touvron et al., 2023), as well as several others.

All transformer-based models are trained to optimize predictive accuracy. Put simply, during training, the model is presented with a sequence of tokens (the context) with one token missing. As we will discuss below in Section 3.4, different models vary in the type of context they use during training. However, in all such models, the task is to predict the missing token based on the context provided. Over the course of training, the model adjusts its parameters (including weights in "attention" mechanisms that determine the relative importance of relationships between tokens) to approximate the conditional probability distribution of tokens that could occur, given the context provided. This enables the model to estimate the probability of each candidate token. In addition to this type of unsupervised pre-training, models may undergo supervised (or unsupervised) "fine-tuning" to further optimize their parameters for any particular task or purpose.

Once fully trained, a model can take any context as an input and generate probabilities for any target token given that context. These token probabilities can then be converted into lexical probabilities (see Section 4.1 below for details on this conversion process).

[7] GPT refers to the underlying architecture. The well-known ChatGPT is a specific implementation of GPT, fine-tuned for conversational tasks, such as responding to prompts in a chatbot-like format.

3.4 Why we used GPT-3 'davinci_002' to estimate lexical probability values in the present study

As noted above, there are several transformer-based LLMs that are able to successfully learn normative patterns of language. However, these models are trained in different ways. Some LLMs, like BERT, are trained to predict a "masked" token by using *bilateral* context—i.e., during training, *both* the preceding and following tokens in a sequence are considered (Devlin et al., 2018). This method, called masked language modeling (MLM), can be effective. However, it differs from how humans process language during real-time communication, where only the *prior* context is available to the comprehender at any given time.

For the present study, we chose to use GPT-3. The GPT series of models are trained to predict the next token based solely on *prior* ("left") context. This was particularly important for our study because we were specifically interested in estimates of lexical probability that rely exclusively on the preceding context. This is because preceding context is the *only* information comprehenders have access to during real-time speech processing; to maintain coherence, producers must therefore be sensitive to lexical probability based on the same context that is available to the comprehender. This is why estimates from GPT (and other left-context models, such as LLaMa) are frequently used in large-scale psycholinguistic studies. As noted above, these studies show that probability estimates from these models are strong predictors of both behavioral response times (e.g. Shain, Meister, Pimentel, Cotterell, & Levy, 2024; Boyce & Levy, 2023), and evoked neural responses during natural language processing (e.g., Michaelov et al., 2024; Szewczyk & Federmeier, 2022).

At the time of data collection, GPT-3 was the most advanced iteration of this model

series, making it the ideal choice for our analyses. Within the GPT-3 family, there are different model variants, some of which undergo additional supervised fine-tuning to adapt the model to specific tasks. For this study, we chose to use the 'davinci-002' variant, described by OpenAI as an improvement over the base 'davinci' model. Davinci-002 is fine-tuned to optimize acceptability of responses across a *wide variety* of tasks and instructions. This makes it particularly flexible and capable of responding coherently to long, complex contexts. Although fine-tuning could, in theory, introduce biases or unexpected effects, we believe that this variant is well-suited to our study's design, in which participants engage in a structured task (picture description) following verbal instructions. We therefore believe that the fine-tuning applied to davinci-002 would further optimize its ability to provide accurate estimates of lexical probability in our speech data.

3.5 What one can and cannot infer from using LLMs to estimate lexical probability values

It is important to emphasize that there is much we do not yet know about GPT and other large language models (LLMs). The primary limitation of these models is that it is usually unclear *how* they generate their probability distributions: Neural networks are often described as "black box" methods because the parameters and representations they learn are not inherently interpretable. As a result, their internal representations may not map straightforwardly onto the representations learned by humans. In addition, most current LLMs cannot tell us exactly *how* higher-level representations influence lower-level lexical information during language processing because their architectures are feedforward in nature. In contrast, the human brain is characterized by long-range feedback connections that bridge the highest and lowest levels of the

cortical hierarchy and are deeply integrated within the cortical microcircuitry . We consider both these limitations in the Discussion section of the main manuscript. We also note that we know relatively little about the precise training data provided to GPT-3, or the specific procedures used for fine-tuning the 'davinci-002' variant.

We do, however, know the following:

(a) GPT-3's training corpus is extremely large (45 TB of text data)—large enough for the model to capture a substantial portion of the statistical variation that characterizes natural language output;

(b) GPT-3 has been highly successful in translating these statistical regularities into probabilistic predictions about upcoming tokens, as evidenced by its ability to generate coherent and fluent output (see, e.g., Brown et al., 2020 for performance on standardized benchmarks);

(c) There are close correspondences between GPT outputs and the probabilities of words estimated more directly using Cloze probability (e.g., Sharpe, 2023; Hao, Mendelsohn, Sterneck, Martinez, & Frank, 2020; Szewczyk & Federmeier, 2022).

(d) As discussed in the previous section, GPT-estimated probability values are strong predictors of human behavioral responses (e.g. Shain, Meister, Pimentel, Cotterell, & Levy, 2024; Boyce & Levy, 2023), and neural activity (e.g., Michaelov et al., 2024; Szewczyk & Federmeier, 2022).

For all these reasons, we can be confident that GPT-3 effectively captures contextual

dependencies that are critical for producing coherent and normatively structured speech. It is this capacity to accurately represent contextual dependencies that we exploit in the present study.

4. The Present Study: Extraction of Lexical Probability Values Under Multiple Conditions

4.1 Estimation of lexical probability values from GPT-3

The initial transcripts of patients' speech were prepared for the purpose of clinical ratings of thought disorder (see Supplement 1.2 and 1.3). Several preprocessing steps were then performed on these transcripts before obtaining probability values from GPT-3 'davinci_002'. First, each participant's transcript was divided into contiguous speech "segments," defined as utterances uninterrupted by experimenter speech. Second, all spellings were manually standardized. Third, punctuation was automatically removed, except for apostrophes, hyphens, and sentence-final punctuation. Sentence-final words were marked with a period or, in the rare case of an obvious question, with a question mark (e.g., "What is that?"). The proportion of sentence-final words (i.e., those containing sentence-final punctuation) to total words was generally low (approximately 0.07) and did not differ significantly between patients and controls.

We accessed OpenAI's GPT-3 'davinci_002' through the OpenAI API (https://openai.com/product) and entered each segment into the model as a string of text. Using a custom-written script, we extracted the probability of each token based on the set of tokens that preceded it.

The probability values we obtained represent GPT's estimate of the token's probability in relation to all possible continuations (i.e. we did not truncate the distribution of possible continuations). For words that were made up of more than one token, e.g. "grasshopper" (which is composed of tokens "grass", "hop", and "per"), we computed lexical probability using the rule of joint probabilities, as follows:

$P(W \mid Context) = P(t1 \mid Context) * P(t2 \mid Context + t1) *, \ldots, * P(tn \mid Context + t1 + t2 +, \ldots, + tn\text{-}1),$ where the word W is made up of n tokens t1, t2, ..., tn.

Because GPT-3 is hosted remotely by OpenAI, and because our study used human subjects' speech data, privacy was a potential concern for data collection. We took a number of steps to mitigate this concern. First, all subject data was anonymized, and all speech samples were checked to ensure they contained no identifying information. Second, we opted out of allowing OpenAI to use our data to improve their models. Finally, after data was collected, the OpenAI account used for data collection was deleted; per OpenAI's policies at the time of data collection, this required OpenAI to delete any and all data associated with the account. Thus, no identifiable data was input into GPT-3, and no data was stored long-term on OpenAI's servers.

Since the time of data collection, OpenAI has removed the ability to extract probability values from GPT-3. However, GPT-2, an earlier version of GPT-3, can still be used to obtain predictability values (for code to interact with a pre-trained GPT-2 model, see https://github.com/samer-noureddine/GPT-2-for-Psycholinguistic-Applications). Values obtained

via GPT-2 are closely correlated with those obtained from GPT-3 (for a subset of our data, Pearson's *r* was approximately .9).

We also note that recent open-source variants of GPT (e.g. GPT-Neo; Black, Gao, Wang, Leahy & Biderman, 2021, as well as other models such as LLaMa (Touvron et al., 2023), *do* allow users to extract probability values. These models, like GPT-3, are trained to predict tokens only based on *prior* context; therefore, they provide alternatives for generating the measures we describe in the present study.

As described in the main manuscript, we obtained multiple different estimates of lexical probability for each word based on varying types and lengths of context. Below, we walk the reader through the process of obtaining these values.

4.2 Extraction of lexical probability values based on all available context

For our first analysis (see Results, Section 1), we were interested in testing the broad hypothesis that patients with first-episode psychosis were generally less sensitive to context than healthy controls. We therefore computed a measure of the lexical probability of every word within each speech segment, given on all the words preceding that word, i.e., P(word | AllAvailableContext).

For example, imagine that a participant produced the following speech segment:

"This is an old picture of a girl in a garden. She's sitting down next to a very pretty rose bush, and on one of the leaves she sees this grasshopper with really shiny wings."

From the second word onward (because probability values cannot be obtained for tokens with no context), we obtained the probability of each token in the string given all the tokens that came before it. Thus, the lexical probability of "is" given all available context, would be P(is | This); the lexical probability of "an" given all available context would be P(an | This is); the lexical probability of "old" given all available context would be P(old | This is an). And so on for every word in every speech segment.

As noted above, if multiple tokens corresponded to a single word, their values were combined into a single probability value using the rule of joint probabilities. Thus, for the above example, we obtained the following lexical probabilities, where each value represents the probability of the word occurring given the entire context that preceded it:

Word	**Lexical Probability**
This	-
is	0.1653
an	0.1946
old	0.0000

picture	0.0470
of	0.7021
a	0.2376
girl	0.0102
in	0.1083
a	0.3432
…	…

Table 1. Lexical probabilities in the AllAvailableContext position. Probabilities represent the probability of Word n *given all the words (up to* n-1*) before it, where* n *is the index/position of the word within the speech segment (1, 2, 3, and so on).*

4.3 Extraction of lexical probability based on unrelated scrambled contexts (baseline)

Also for the purpose of our first analysis (Results, Section 1), we sought a "baseline" probability measure that would reflect the probability of each word in each speech segment, given the same number of preceding words of context, but in the absence of any type of contextual structure or related content (i.e. without the influence of semantic, syntactic, discourse-level, or other dependencies). This allowed us to rule out the possibility that any between-group differences in P(word | AllAvailableContext) were driven simply by differences in the types of words patients and controls used.

We extracted this baseline measure of probability for each word in each speech segment by providing GPT-3 with that target word preceded by the same number of words that preceded it in the original speech sample, but with these context words selected randomly from a description of an unrelated picture, i.e. P(Word | UnrelatedScrambledContext). This unrelated picture description was simulated using ChatGPT (see Section 1.4 of this Supplement for the instructions we gave to ChatGPT to obtain this speech sample, as well as the simulated speech sample itself).

For example, returning to the example speech segment introduced above (Section 4.2), to obtain the baseline probability for the third word in the segment ("an"), instead giving the true context to GPT-3 ("This is"), we might give it two randomly selected words from the unrelated simulated speech sample (e.g. "violin to"), and so we would extract the probability P(an | violin to). Similarly, to obtain the probability of the fourth word ("old"), instead of inputting "This is an " into GPT-3 as the context, we might input the string, "so got sleeves", thus extracting the probability P(old | so got sleeves). In this way, for each target word in each speech segment, we generated a value for P(Word | UnrelatedScrambledContext). A unique scrambled context was generated for each target. Unsurprisingly, these scrambled context lexical probabilities were all very small.[8]

[8] Note that the specific text used to generate this scrambled unrelated context is not critical; the purpose was, simply, to generate contexts with no coherent relationship to the upcoming words (i.e. to disturb all types of dependencies).

4.4 Extraction of lexical probability based on context windows of varying length (1-50 words)

A major goal of the present study was not only to determine whether there was an overall difference between patients and controls in the probability of words they produce, given all available context, but also to determine whether this difference is influenced by the length of prior context (number of words). Specifically, we were interested in whether positive thought disorder is characterized by a relative insensitivity to global vs. local context.

Therefore, for each word in each speech sample, we extracted multiple measures of lexical probability, based on multiple different context window sizes (with context window size defined as number of words in the prior context). Specifically, for each word in each participant's speech segment, we systematically titrated the length of the context provided to GPT-3, from very local (1 word of preceding context) to more global (up to 50 words of preceding context).

Consider once again the example speech segment described above (repeated below for convenience):

"This is an old picture of a girl in a garden. She's sitting down next to a very pretty rose bush, and on one of the leaves she sees this grasshopper with really shiny wings."

Here, the word "leaves" is preceded by 26 words. So, for this word, we would use GPT-3 to

extract 26 different probability values (given 1 word of context: P(leaves | the); given 2 words of context: P(leaves | of the) given 3 words of context: P(leaves | one of the), given 4 words of context: P(leaves | and one of the) and so on up to 26 words of context.

For the 51st word of each speech segment, and all words following it, we obtained 50 probability values: the probability of the target word given a single word of context, its probability given 2 words of context, and so on up to 50 words of context.

Obviously, it is not possible to extract probability given a context window size longer than the length of the *available* context (e.g. we cannot have a 40-word context window for the 35th word in a speech segment). Therefore, target words 1 through 50 of each speech segment had missing values for any context window sizes greater than n-1 (where n is the position of the word — 1st, 2nd, etc.— in the speech segment). Note that linear mixed effects analyses are well-equipped to handle uneven numbers of datapoints across levels of a predictor, and so this did not pose problems for our analyses.

Extracting the probability of each word, for each context window size, yielded a matrix that looked like this:

Word	**1Word_Context Window**	**2Word_Context Window**	**3Word_Context Window**	**…**
This	-	-	-	…
is	0.1653	-	-	…
an	0.0093	0.1946	-	…
old	0.0000	0.0000	0.0000	…
picture	0.0081	0.0322	0.0452	…
of	0.0282	0.3044	0.5033	…
a	0.0056	0.1555	0.2123	…
girl	0.0000	0.0021	0.0063	…
in	0.0122	0.0488	0.0359	…
a	0.0020	0.2045	0.1923	…
garden.	0.0000	0.0000	0.0000	…
She's	0.0003	0.0004	0.0005	…
…	…	…	…	

Table 2. Lexical probabilities given a range of ContextWindowSizes (1-3 shown). Cells filled in with dashes (-) represent missing values, where it was impossible to compute probability for a given word, given the context window size.

5. Preparation of Data for Statistical analysis

5.1 Subsequent exclusion of function words, disfluencies, and sentence-final punctuation prior to analysis

As described above, for the purposes of obtaining lexical probability values from GPT-3, we entered the full transcribed speech segments as strings of text — content words, function words and disfluencies, and sentence-final punctuation and any punctuation included in the spelling of particular words.

However, for the purpose of statistical analysis, we subsequently excluded the probability values for function words (based on the stopwords list from the Natural Language Toolkit; Bird, Klein, & Loper, 2009) and disfluencies. This therefore left, for analysis, only datapoints corresponding to the probability of *content words*. The reason for these exclusions is that we did not expect the probability values of function words or disfluencies to be heavily influenced by global sources of information in *either* group. In contrast, we know that, in healthy adults, the probabilities of content words are influenced by both local and global sources of information (see Supplement, 2.1 and 2.3 above for elaboration). Therefore, restricting our analyses to probability values of content words allowed us to test our primary hypothesis of interest: that patients' sensitivity to global (vs. local) context is disproportionately reduced.

5.2 Final formatting of data for statistical analysis

As described in the main manuscript, our main question of interest (see Results, Section 2)

involved testing for an interaction between Context Window Size (ranging from 1 to 50; log-transformed) and Group (patients vs. controls) on lexical predictability. This required us to convert the matrix shown above (Section 4.4) to "long" format, for use in our linear mixed effects models (LMERs).

Word	**ContextWindowSize**	**Lexical Probability**
old	1	0.0000
old	2	0.0000
old	3	0.0000
picture	1	0.0081
picture	2	0.0322
picture	3	0.0452
…	…	…

Table 3. Lexical probability data in "long" format for analysis. All words except content words have been excluded from the dataset.

6. Assumptions for Hypothesis Testing and Ruling Out Artifactual Explanations for the Group*ContextWindowSize Interaction

Linear mixed effects regression (LMER) has emerged as a powerful tool in psychology research, in part due to its ability to deal with clustering in the data at multiple levels, while allowing the researcher to include "nuisance" covariates to control for potential confounds. However, like all analytic approaches, LMER was developed with certain assumptions in mind. Here, we first outline four relevant statistical assumptions for linear regression and explain how we checked these in the present study. We then discuss why the observed Group × ContextWindowSize interaction cannot easily be attributed to a statistical artifact, arising either from scaling-related distortions (e.g., a failure to apply appropriate log transformations, or a floor effect), or from a more general impairment in the patient group (e.g., a more stochastic generative process or increased trial-by-trial variability).

6.1 Verifying statistical assumptions

The following four assumptions are generally recognized as specifying the ideal conditions for linear regressions to ensure that the results are interpretable and that p-values are not inflated.

(a) Independent variables are related linearly to the dependent variable.

To ensure a linear relationship between, we log transformed the probability values for each word (our dependent variable) as well as the independent predictor variable ContextWindowSize. We were quite confident from previous work that we would see this linear relationship (Sharpe, 2023), and, indeed, plots of our data supported the assumption of linearity in our main interaction models (see, however, discussion of possible non-linearity in our

Disorganization*ContextWindowSize model below in Section 8). These log transformations also eliminated potential non-linear distortions in the relationship between probability and context window size (see Section 6.2(a) below).

(b) The errors, or residuals, have constant variance (homoscedasticity).

Although probability data is often severely heteroscedastic, the log-transformation has the effect of "spreading out" datapoints near zero. This greatly reduced heteroskedasticity in our models, thus bringing us more closely in line with the assumptions of the general linear model. Indeed, diagnostic plots revealed minimal heteroscedasticity associated with our predictors of interest.

(c) The residuals are independent of one another.

In our models, non-independence due to by-participant and by-word clustering was accounted for using a mixed effects model, with random intercepts and random slopes with respect to our predictors of interest. Another potential source of non-independence in data is autocorrelation. However, in our data, exploration of the data clearly showed that lexical probability at word *t* was not significantly correlated with lexical probabilities at words *t-1*, *t-2*, or *t-3*. Examination of an autocorrelation function led us to the same conclusion: that there was little to no autocorrelation impacting our model.

(d) The residuals are normally distributed.

Based on a series of diagnostic plots, we determined that our residuals were not quite normal: symmetric but moderately light-tailed. Given that normality of residuals is thought to be the least important assumption for statistical inference (that is, p-values for linear regression

appear quite robust to violations of this assumption (e.g. Gelman & Hill, 2006; Schmidt & Finan, 2018), and given the other assumptions were met in our data, we had no reason to believe this non-normality affected our results. However, out of an abundance of caution, we carried out additional analyses which did not assume a normal distribution of residuals (see Section 8.1(d) below).

6.2 Ruling out alternate explanations for the Group by ContextWindowSize interaction

A key finding of this study was a significant Group × ContextWindowSize interaction. We interpret this statistical interaction as evidence for a selective insensitivity to more global context, relative to more local context, in the patient group. In this section, we explain why this effect cannot be trivially attributed to non-linear distortions in the relationship between probability and context window size or a floor effect (scaling-related distortions), or from increased stochasticity or increased trial-by-trial variability in the patient group (differences between groups in the underlying generative process or noise).

(a) Non-linear distortions due to compounding effects of context

As we show in the present study, lexical probability in healthy controls increases with increasing context window size. This is because, in general, additional units of context build on the previous context. However, this build-up of context also means that, in terms of raw probability, deficits in using context, at any level, compound as context becomes more and more global. To illustrate this phenomenon, assume that each additional unit of context increases

probability in healthy controls by approximately .1, such that, a word's probability = .1 when preceded by one unit of context, .2 when preceded by two units of context, .3 when preceded by three units of context, and so on. Now assume that patients exhibit a *non-selective* deficit in their use of context, whereby each unit of context, no matter how much context precedes it (i.e. how local or global it is), is approximately 50% as beneficial as in healthy controls (an increase of .05). Thus, in patients, a word's lexical probability = .05 when preceded by one unit of context, .1 when preceded by two units of context, and .15 when preceded by three units of context. Notice that, with each additional unit of context, the absolute difference between patients' and controls' lexical probabilities increases (.05 at one unit of context, .1 at two units of context, .15 at three units of context, and so on), would produce an interaction between Group and Context Window Size. Thus, using untransformed lexical probabilities as the dependent variable in linear regression analysis cannot distinguish a non-selective deficit in context from a selective one.

For this reason, we chose to use log-transformed probabilities — specifically the natural log of raw lexical probabilities — in order to examine *proportional* changes in probabilities with increasing context window size. In the above example, the difference between patients' *log probabilities* does *not* change with increasing context window size: ln(.1) - ln(.05) = .69; ln(.2) - ln(.1) = .69; and ln(.3) - ln(1.5) = .69. The difference remains constant/proportional, and there is therefore no interaction with context window size. In contrast, a *selective* deficit in global context, by which each additional unit of context is increasingly less beneficial to patients' speech, *will* produce an interaction between group and context window size on log lexical probabilities, as the difference between patients' and controls' lexical probabilities becomes disproportionately larger.

Note that our use of the natural log means the parameter estimates from our LMER models can be interpreted, roughly, in terms of percent changes in lexical probability. Take, for example, the main effect of Group on Global Lexical Probability, as shown in Table 3B of the main manuscript. Here, the parameter estimate is -.25. To get the difference in raw lexical probability between patients and controls, exponentiate the absolute value of the coefficient to get approximately 1.28, then subtract 1 and multiply by 100. Global lexical probabilities in patients, therefore, are about 28% smaller than in healthy controls.

The same process can be applied to continuous effects. Take, for example, the main effect of log ContextWindowSize on Lexical Probability, as shown in Table 3A of the main manuscript. The parameter estimate is .81; to get the percent increase in lexical probability for a one unit increase in log ContextWindowSize, exponentiate the coefficient to get approximately 2.25, then subtract 1 and multiply by 100. Lexical probability, therefore, increases 125% for each unit increase in log ContextWindowSize.

(b) A floor effect?

Whenever a Group × Predictor interaction is observed, it is also important to consider the possibility that it reflects a floor effect artifact; that is, it is important to rule out the possibility that, at the lowest end of the predictor range, the effect of the predictor on the dependent measure is difficult to detect in either group.

This explanation seems unlikely in the present study. First, at the lowest end of the context window size scale, as the context window size increased from 1 to 3 words, both groups

exhibited measurable increases in log-transformed lexical probability; that is, in both groups small amounts of contextual information were sufficient increase lexical probability. Second, at these small context window sizes, both groups exhibited log lexical probabilities that exceeded those observed in the UnrelatedScrambledContext baseline condition, indicating that the dependent measure was not compressed against floor. Third, the Group × ContextWindowSize interaction was not driven solely by effects at the smallest context window sizes, as would be expected under a floor effect account. Instead, it reflected a gradual, proportional reduction in patients' sensitivity to increasing context size, with the divergence becoming progressively more pronounced at larger (more global) context window sizes.

(c) More stochastic word production in patients at all window sizes?

We also considered the possibility that patients' word production was generally more random than that of controls, irrespective of context window size. This would not represent a scaling artifact, but rather a mechanistic claim—namely, that patients rely on a more stochastic generative process during language production, i.e., one that is less constrained by prior context and thus less sensitive to changes in contextual information at any window size. This more general account could, however, give rise to an artifactual Group × ContextWindowSize interaction, even if patients were not selectively impaired in their use of global (vs. local) context. Because lexical probability typically increases as a function of context window size, a more stochastic generative process in patients would make it more difficult to detect any effect of context window size in this group, potentially producing a spurious interaction between Group and ContextWindowSize.

In the present study, two pieces of evidence suggest that this explanation cannot easily account for our findings. First, if patients' language production was generally more stochastic, then one would expect some reduction in lexical probability at the smallest context window sizes. However, at these very local context windows, we observed no evidence of reduced lexical probabilities in patients relative to controls and both groups exhibited the same increase in lexical probability as the context window size increased from 1 to 3 words.

Second, if patients' word production was generally more stochastic than that of controls, they should produce less probable words under any type of contextual manipulation, relative to an UnrelatedScrambledContext baseline. However, as reported below in Section 8.2 when we compared participants' ability to use the broad semantic field established by the prior context with this UnrelatedScrambledContext baseline, both groups benefited equally, with no main effect of Group and no significant Group × Condition interaction.

(d) *Increased variability in patients vs. controls?*

Finally, we considered the possibility that the Group × Condition interaction arose because probability values were more variable across trials in the patient than the controls. This explanation differs from the generalized stochasticity account described above because it assumes that patients are capable of using prior context, but intermittently fail to do so due to non-specific factors such as reduced motivation or fluctuating attention. As with the generalized stochasticity account, however, increased trial-to-trial variability in the patient group could make it more difficult to detect an effect of context window size in the patient than the control group, leading to an apparent Group*ContextWindowSize interaction, even in the absence of a true

selective impairment.

This account seems unlikely for the same two reasons described above. In addition, range and variance of log-transformed lexical probability revealed that both were comparable across groups, at both local and global context levels.

7. Complete Results Tables for Analyses Reported in the Main Manuscript

In the main manuscript, the results tables report findings only for the predictors of interest. Here, we report the full results of each LMER, including nuisance covariates. Beneath each table caption, we also include the R code for defining and testing each statistical model.

Effect of Group and ContextType (NoContext vs. AllAvailableContext) on Lexical Probability

	Estimate	SE	t	p	Sig.
Intercept	-6.54	0.04	-158.12	0.00	***
Group	**-0.23**	**0.08**	**-2.72**	**0.01**	**
Context Type	**4.63**	**0.04**	**109.10**	**0.00**	***
PSES	-0.05	0.04	-1.32	0.19	
SegmentLength	0.34	0.02	17.06	0.00	***
ContextType*Group	**-0.29**	**0.09**	**-3.35**	**0.00**	**
ContextType*PSES	-0.02	0.04	-0.45	0.65	
ContextType*SegmentLength	0.17	0.04	4.75	0.00	***

Supplementary Table 4. Predictors of interest are shown in bold.

*Mod1 = LexicalProbability ~ 1 + Group*ContextType + PSES*ContextType + SegmentLength*ContextType + (1 + ContextType | Participant) + (1 + ContextType | Word)*

Lmer1 = lmer(Mod1, data = data, control = lmerControl(optCtrl = list(maxfun = 100000000), optimizer = 'Nelder_Mead'))

Effect of Group and ContextWindowSize on Lexical Probability

	Estimate	SE	t	p	Sig.
Intercept	-4.65	0.04	-106.33	0.00	***
ContextWindowSize	**0.81**	**0.01**	**82.83**	**0.00**	***
Group	**-0.34**	**0.09**	**-3.75**	**0.00**	***
PSES	-0.07	0.04	-1.74	0.09	.
SegmentLength	-0.03	0.03	-1.32	0.19	
ContextWindowSize*Group	**-0.08**	**0.02**	**-4.04**	**0.00**	***
ContextWindowSize*PSES	-0.01	0.01	-0.61	0.54	
ContextWindowSize*SegmentLength	-0.03	0.01	-4.20	0.00	***

Supplementary Table 5. Predictors of interest are shown in bold.

*Mod2 = LexicalProbability ~ 1 + Group*ContextWindowSize+ PSES* ContextWindowSize + SegmentLength* ContextWindowSize + (1 + ContextWindowSize | Participant) + (1 + ContextWindowSize | Word)*

Lmer2 = lmer(Mod2, data = data, control = lmerControl(optCtrl = list(maxfun = 1000000), optimizer = 'bobyqa'))

Effect of Overall Cognition, Group, and ContextWindowSize on Lexical Probability

	Estimate	SE	t	p	Sig.
Intercept	-4.50	0.08	-58.57	0.00	***
CognitiveFunction	**0.01**	**0.06**	**0.21**	**0.83**	
ContextWindowSize	**0.84**	**0.02**	**48.11**	**0.00**	***
Group	**-0.31**	**0.11**	**-2.78**	**0.01**	**
PSES	-0.09	0.05	-1.73	0.09	.
SegmentLength	-0.04	0.03	-1.21	0.23	
ContextWindowSize*CognitiveFunction	**0.00**	**0.01**	**-0.26**	**0.80**	
ContextWindowSize*Group	**-0.08**	**0.03**	**-3.24**	**0.00**	**
ContextWindowSize*PSES	-0.01	0.01	-0.89	0.38	
ContextWindowSize*SegmentLength	-0.03	0.01	-3.39	0.00	***

Supplementary Table 6. Predictors of interest are shown in bold.

*Mod3 = LexicalProbability ~ 1 + Group*ContextWindowSize+ CognitiveFunction*ContextWindowSize +PSES* ContextWindowSize + SegmentLength* ContextWindowSize + (1 + ContextWindowSize | Participant) + (1 + ContextWindowSize | Word)*

Lmer3 = lmer(Mod3, data = data, control = lmerControl(optCtrl = list(maxfun = 100000000), optimizer = 'bobyqa'))

Effect of TLI Disorganization and ContextWindowSize on Lexical Probability (Within Patients)

	Estimate	SE	t	p	Sig.
Intercept	-4.79	0.06	-79.74	0.00	***
Disorganization	**-0.08**	**0.04**	**-2.00**	**0.05**	**.**
ContextWindowSize	**0.78**	**0.01**	**59.00**	**0.00**	***
Segment_Length	-0.05	0.04	-1.47	0.14	
PSES	-0.10	0.06	-1.82	0.07	.
ContextWindowSize*Disorganization	**-0.02**	**0.01**	**-2.75**	**0.01**	**
ContextWindowSize*Segment_Length	-0.03	0.01	-3.28	0.00	**
ContextWindowSize*PSES	0.01	0.01	0.54	0.59	

Supplementary Table 7. Predictors of interest are shown in bold.

*Mod4 = LexicalProbability ~ 1 + Disorganization*ContextWindowSize+ CognitiveFunction*ContextWindowSize +PSES* ContextWindowSize + SegmentLength* ContextWindowSize + (1 + ContextWindowSize | Participant) + (1 + ContextWindowSize | Word)*

Lmer4 = lmer(Mod4, data = data[which(data$Group == "FEP"),], control = lmerControl(optCtrl = list(maxfun = 100000000), optimizer = 'bobyqa'))

Effect of TLI Impoverishment and ContextWindowSize on Lexical Probability (Within Patients)

	Estimate	SE	t	p	Sig.
Intercept	-4.81	0.06	-83.12	0.00	***
Impoverishment	**-0.09**	**0.05**	**-1.77**	**0.08**	**.**
ContextWindowSize	**0.77**	**0.01**	**57.84**	**0.00**	***
Segment_Length	-0.07	0.04	-1.89	0.06	**.**
PSES	-0.10	0.06	1.81	0.08	.
ContextWindowSize* Impoverishment	**0.00**	**0.01**	**-0.35**	**0.73**	
ContextWindowSize*Segment_Length	0.03	0.01	-3.54	0.00	***
ContextWindowSize*PSES	0.01	0.01	0.54	0.59	

Supplementary Table 8. Predictors of interest are shown in bold.

*Mod5 = LexicalProbability ~ 1 + Impoverishment*ContextWindowSize+ CognitiveFunction*ContextWindowSize +PSES* ContextWindowSize + SegmentLength* ContextWindowSize + (1 + ContextWindowSize | Participant) + (1 + ContextWindowSize | Word)*

Lmer5 = lmer(Mod5, data = data[which(data$Group == "FEP"),], control = lmerControl(optCtrl = list(maxfun = 100000000), optimizer = 'bobyqa'))

Effect of TLI Disorganization and ContextWindowSize on Lexical Probability Over and Above PANSS-8 (Within Patients)

	Estimate	SE	t	p	Sig.
(Intercept)	-4.86	0.08	-60.85	0.00	***
Disorganization	**-0.07**	**0.04**	**-1.70**	**0.10**	**.**
ContextWindowSize	**0.78**	**0.02**	**44.42**	**0.00**	***
PSES	-0.10	0.06	-1.75	0.09	.
SegmentLength	-0.04	0.04	-1.22	0.22	
PANSS-8	**0.07**	**0.08**	**0.93**	**0.36**	
Disorganization*ContextWindowSize	**-0.02**	**0.01**	**-2.80**	**0.01**	**
ContextWindowSize*PSES	0.01	0.01	0.92	0.36	
ContextWindowSize*SegmentLength	-0.03	0.01	-3.25	0.00	**
ContextWindowSize*PANSS-8	**0.00**	**0.02**	**0.15**	**0.88**	

Supplementary Table 9. Predictors of interest are shown in bold.

*Mod6 = LexicalProbability ~ 1 + Disorganization*ContextWindowSize+PANSS8*ContextWindowSize + CognitiveFunction*ContextWindowSize +PSES* ContextWindowSize + SegmentLength* ContextWindowSize + (1 + ContextWindowSize | Participant) + (1 + ContextWindowSize | Word)*

Lmer6 = lmer(Mod6, data = data[which(data$Group == "FEP"),])

8. Additional Statistical Analyses

8.1. Robustness of main statistical results reported in main manuscript

To test the robustness of the findings reported in the main manuscript, we carried out a number of exploratory analyses on the main interaction models — Group*ContextType (Results, Section 1), Group*ContextWindowSize (Results, Section 2), Disorganization*ContextWindowSize (Results, Section 4), and Impoverishment*ContextWindowSize (Results, Section 4).

(a) *Repetition of each analysis without nuisance covariates*

As noted in the Methods, we included Parental SES and Segment Length (in words) as nuisance variables. Following recommendations of Becker (2005) and Kalnins (2018), we followed up each analysis by repeating the same analysis *without* these nuisance covariates to exclude the possibility of spurious "suppression" effects which can occur when predictors are colinear. All significant interactions remained significant after removing covariates, while the Impoverishment*ContextWindowSize interaction remained non-significant.

(b) *Analyses including additional nuisance covariates*

We elected *a priori* to include Parental SES and Segment Length (in words) as nuisance variables for our main analyses. In addition, we also carried out post-hoc analyses using other nuisance covariates: ParticipantAge (in years), WordOrderInSegment (numbered 1 through the length of the segment), ProportionUniqueWords (encodes number of unique words produced by

an individual, divided by the total words produced), SentenceLength (in words), MedianSentenceLength (over the entire speech segment), 1BackRepetition (i.e., whether a word *t* is a repetition of word *t-1*,e.g. “dog dog”), and 2BackRepetition (i.e. whether a word *t* is a repetition of word *t-2*, e.g. “dog the dog”). The pattern of results remained the same after inclusion of these additional covariates.

(c) *Robustness to different outlier removal methods and different data transformations*

Our findings were robust to different outlier removal methods (standard deviation cutoff versus quantile trimming) and thresholds (2, 2.5, or 3 SDs from the mean), as well as to different data transformations (e.g. logit).

(d) *Beta regression to account for non-normality of residuals*

For all analyses we used log-transformed probability values as our dependent measure. This had the effect of bringing our model largely in line with the assumptions of linear mixed effects regression (see Section 6.1(a)) and also allowed us to capture proportional changes in lexical probability with increasing context window size (see Section 6.2(a)).

However, because our residuals were not quite normally distributed (see Section 6.1(d)), out of an abundance of caution, for the main interaction models, we conducted additional analyses in which we treated the dependent variable as raw (untransformed) probabilities and modeled it as a beta-distributed outcome variable. This approach is well suited for modeling

dependent variables that represent probabilities, as it assumes the data follow a beta distribution—a continuous probability distribution constrained between 0 and 1. In all cases, we saw the same pattern of results that we report in the main manuscript.

In the table below, we present the results of one of these analyses — the analysis that tested for a Group*ContextWindowSize interaction. Note that due to convergence issues when both by-participant and by-items random effects were specified, we ran two versions of this model — one accounting for by-participant random effects, and one accounting for by-item random effects. The pattern of results for predictors of interest is the same in both.

Effect of ContextWindowSize*Group on Lexical Probability (Beta Model; By-Participant Random Effects)

	Estimate	SE	z	p	Sig.
Intercept	-1.74	0.00	-825.54	0.00	***
Group	-0.06	0.00	-21.52	0.00	***
ContextWindowSize	**0.21**	**0.00**	**61.42**	**0.00**	***
ParentalSES[a]	-0.01	0.00	-3.85	0.00	***
SegmentLength	0.02	0.00	11.36	0.00	***
Group*ContextWindowSize	**-0.03**	**0.01**	**-3.92**	**0.00**	***
ContextWindowSize*ParentalSES	0.00	0.00	0.21	0.83	
ContextWindowSize*SegmentLength	0.01	0.00	2.99	0.00	***

Supplementary Table 10 . Predictors of interest are shown in bold. LexicalProbability was not log-transformed.
[a]ParentalSES = Parental Socioeconomic Status.

Model Structure: LexicalProbability~ Group*ContextWindowSize + ParentalSES*ContextWindowSize + SegmentLength*ContextWindowSize + (0 + ContextWindowSize||Participant)

Effect of ContextWindowSize*Group on Lexical Probability (Beta Model; By-Word Random Effects)

	Estimate	SE	z	p	Sig.
Intercept	-1.85	0.03	-65.38	0.00	***
Group	-0.05	0.04	-1.29	0.20	
ContextWindowSize	**0.84**	**0.01**	**111.70**	**0.00**	***
ParentalSES[a]	-0.01	0.02	-0.71	0.48	
SegmentLength	0.07	0.02	3.63	0.00	***
Group*ContextWindowSize	**-0.04**	**0.01**	**-3.69**	**0.00**	***
ContextWindowSize*ParentalSES	-0.01	0.01	-1.03	0.30	
ContextWindowSize*SegmentLength	-0.01	0.01	-0.96	0.34	

Supplementary Table 11. Predictors of interest are shown in bold. LexicalProbabilitywas not log-transformed.
[a]ParentalSES = Parental Socioeconomic Status.

Model Structure: LexicalProbability~ Group*ContextWindowSize + ParentalSES*ContextWindowSize + SegmentLength*ContextWindowSize + (1 + ContextWindowSize|Word)

8.2 Are patients able to benefit from the broad semantic relationships between each word and the preceding set of words they had produced?

As described in Section 1 of the Results, we first asked whether patients were *generally* less sensitive to context than healthy controls. To do this, we extracted the lexical probability of each word in each speech segment, based on the full set of words the participant had produced up to that point, i.e., P(Word | AllAvailableContext). As a baseline, for each word, we computed P(Word | UnrelatedScrambledContext) by extracting the probability of that same word, given a randomly scrambled set of words, matched in number, drawn from a simulated description of an unrelated picture (see Supplement, Sections 1.4 and 4.3 for details). This baseline allowed us to rule out the possibility that group differences in lexical probability reflected differences in the inherent probabilities of the individual words produced by patients.

Here we report a second analysis that asked whether patients, like controls, were able to

benefit from the broad semantic relationships between each word and the preceding set of words they had produced; that is, whether they were able to benefit from the broad semantic field established by their own prior speech, independent of its structure or organization. For this analysis, we replaced the preceding context with the same words that the participant had produced, but in a randomly scrambled order, and then extracted the probability of each word given this scrambled context: P(Word | RelatedScrambledContext). This approach preserved the broad semantic content of the original context while disrupting its structure. We compared this condition to the UnrelatedScrambledContext condition (where, as described above, the scrambled context was taken from the simulated description of a different unrelated picture),

As expected, lexical probabilities in the RelatedScrambledContext condition were significantly greater than in the UnrelatedScrambledContext condition across both groups (main effect of ContextType). Critically, there was no main effect of Group and no Group × Condition interaction.

Effect of ContextType (ScrambledRelatedContext vs. ScrambledUnrelatedContext)*Group on Lexical Probability

	Estimate	SE	t	p	Sig.
Intercept	-8.38	0.04	-19:	0.00	***
Group	**-0.07**	**0.09**	**-0.75**	**0.46**	
ContextType	**-0.95**	**0.03**	**-33.21**	**0.00**	***
ParentalSES[a]	-0.04	0.04	-0.97	0.33	
SegmentLength	0.20	0.02	10.55	0.00	***
Group*ContextType	**-0.04**	**0.06**	**-0.68**	**0.50**	
ParentalSES*ContextType	0.00	0.03	-0.03	0.98	
SegmentLength*ContextType	0.12	0.03	4.08	0.00	***

Supplementary Table 12 . Predictors of interest are shown in bold. [a]ParentalSES = Parental Socioeconomic Status.

Model Structure: LexicalProbability~ Group*ContextType + ParentalSES*ContextType + SegmentLength*ContextType + (1 + ContextType | Participant) + (0 + ContextType | Word)

Looking at the effect of Group *within* each context type, there was, again, no main effect of Group in the UnrelatedScrambledContext condition, and also no main effect of Group in the RelatedScrambledContext condition.

Effect of Group on Lexical Probability within ScrambledRelatedContext

	Estimate	SE	t	p	Sig.
Intercept	-8.85	0.04	-198.80	0.00	***
Group	**-0.07**	**0.09**	**-0.79**	**0.43**	
ParentalSES[a]	-0.04	0.04	-1.01	0.32	
SegmentLength	0.26	0.02	10.68	0.00	***

Supplementary Table 13 . Predictors of interest are shown in bold. [a]ParentalSES = Parental Socioeconomic Status.

Model Structure: LexicalProbability~ Group + ParentalSES + SegmentLength + (1 + ContextType | Participant)

Effect of Group on Lexical Probability within ScrambledUnrelatedContext

	Estimate	SE	t	p	Sig.
Intercept	-7.90	0.05	-174.92	0.00	***
Group	**-0.06**	**0.09**	**-0.67**	**0.50**	
ParentalSES[a]	-0.04	0.04	-0.83	0.41	
SegmentLength	0.12	0.03	4.58	0.00	***

Supplementary Table 14 . Predictors of interest are shown in bold. [a]ParentalSES = Parental Socioeconomic Status.

Model Structure: LexicalProbability~ Group + ParentalSES + SegmentLength + (1 + ContextType | Participant)

These findings have two important implications.

1. They suggest that the group differences observed in the AllAvailableContext condition (Results, Section 1) are unlikely to reflect a failure of patients to establish and use a broad

semantic field to support word production. Instead, the results point to a more specific impairment in how content is organized or structured. This interpretation aligns with clinical accounts of positive thought disorder, which emphasize disturbances in the organization and flow of speech rather than abnormalities in semantic content *per se*.

2. They provide evidence against the possibility that the Group by ContextWindowSize interaction — the key finding of the present study — arose artifactually either as a consequence of a more stochastic generative process in patients, or increased trial-by-trial variability in the patient group due to reduced motivation or randomly-fluctuating attention. As explained above in Sections 6.2(c) and (d) of this Supplement, both these accounts could potentially make it difficult to detect an effect of context window size in the patient group, producing an apparent Group × ContextWindowSize interaction, even in the absence of a true selective impairment. However, both these explanations would also predict that patients should benefit less than controls from *any* aspect of the prior context. The present finding that patients *were* able to benefit from the broad semantic field of their own prior context to the same extent as controls, relative to a significantly less probable unrelated scrambled baseline, argues against this non-selective account.

8.3. Exploratory testing for Group*ContextWindowSize interaction for each individual picture description

For the analyses reported in the main manuscript, we used speech segments from *all three* pictures that each participant viewed. To begin to evaluate our measure's performance on *subsets* of the dataset, we carried out exploratory analyses which tested for the

Group*ContextWindowSize interaction using the descriptions for each *individual* picture.

Below, we show that the interaction effect remains significant when only data from Picture 1 is analyzed, as well as when only data from Picture 3 is analyzed. The interaction does not reach significance when only data from Picture 2 is analyzed, although the parameter estimate has the same directionality (negative, indicating a smaller increase in lexical probability for patients as context length increases), as in the other analyses. Due to the post-hoc nature of these analyses, we do not want to make too much of these findings. However, it will be interesting for future studies to explore what characteristics of a given picture are best for detecting atypical sensitivity to context.

Effect of ContextWindowSize*Group on Lexical Probability: Picture 1

	Estimate	SE	t	p	Sig.
Intercept	-4.33	0.05	-79.70	0.00	***
ContextWindowSize	**0.81**	**0.01**	**66.32**	**0.00**	***
Group	-0.25	0.11	-2.27	0.03	*
ParentalSES[a]	-0.05	0.05	-1.00	0.32	
SegmentLength	-0.04	0.04	-0.93	0.35	
ContextWindowSize*Group	**-0.08**	**0.02**	**-3.37**	**0.00**	**
ContextWindowSize*ParentalSES	-0.01	0.01	-0.79	0.43	
ContextWindowSize*SegmentLength	-0.05	0.01	-4.97	0.00	***

Supplementary Table 15. Predictors of interest are shown in bold. Includes only speech samples in response to Picture 1. [a]ParentalSES = Parental Socioeconomic Status.

Model Structure: LexicalProbability~ ContextWindowSize*Group + ContextWindowSize*ParentalSES + ContextWindowSize*SegmentLength + (1 + ContextWindowSize | Participant) + (1 + ContextWindowSize | Word)

Effect of ContextWindowSize*Group on Lexical Probability: Picture 2

	Estimate	SE	t	p	Sig.
Intercept	-4.69	0.05	-103.71	0.00	***
ContextWindowSize	**0.80**	**0.01**	**55.96**	**0.00**	***
Group	-0.21	0.09	-2.23	0.03	*
ParentalSES[a]	-0.04	0.04	-0.86	0.39	
SegmentLength	0.02	0.05	0.53	0.59	
ContextWindowSize*Group	**-0.03**	**0.03**	**-1.08**	**0.29**	
ContextWindowSize*ParentalSES	0.00	0.01	0.30	0.77	
ContextWindowSize*SegmentLength	-0.02	0.01	-1.52	0.13	

Supplementary Table 16. Predictors of interest are shown in bold. Includes only speech samples in response to Picture 2. [a]ParentalSES = Parental Socioeconomic Status.

Model Structure: LexicalProbability~ ContextWindowSize*Group + ContextWindowSize*ParentalSES + ContextWindowSize*SegmentLength + (1 + ContextWindowSize | Participant) + (1 + ContextWindowSize | Word)

Effect of ContextWindowSize*Group on Lexical Probability: Picture 3

	Estimate	SE	t	p	Sig.
Intercept	-4.65	0.05	-91.76	0.00	***
ContextWindowSize	**0.81**	**0.01**	**61.70**	**0.00**	***
Group	-0.29	0.10	-2.76	0.01	**
ParentalSES[a]	-0.06	0.05	-1.19	0.24	
SegmentLength	0.04	0.05	0.84	0.40	
ContextWindowSize*Group	**-0.09**	**0.03**	**-3.34**	**0.00**	**
ContextWindowSize*ParentalSES	-0.01	0.01	-1.08	0.29	
ContextWindowSize*SegmentLength	-0.01	0.01	-1.10	0.27	

Supplementary Table 17. Predictors of interest are shown in bold. Includes only speech samples in response to Picture 3. [a]ParentalSES = Parental Socioeconomic Status.

Model Structure: LexicalProbability~ ContextWindowSize*Group + ContextWindowSize*ParentalSES + ContextWindowSize*SegmentLength + (1 + ContextWindowSize | Participant) + (1 + ContextWindowSize | Word)

8.4 Exploratory testing for Group*ContextWindowSize interaction, considering each cognitive score separately

In the main manuscript, we show that an aggregate cognitive score—the mean of three scaled scores from the Digit-Symbol Substitution Test, the Trail-Making Test (Part B), and the Category Fluency Task—could not explain Group differences in the effect of ContextWindowSize on lexical probability. Here, we re-run this analysis with each cognitive score separately. None of these three scores interacted significantly with ContextWindowSize. However, in all analyses, the Group*ContextWindowSize interaction persisted. This further supports the idea that our findings cannot be explained by global differences in cognitive functioning between groups.

Effects of DSST Score, Group, and ContextWindowSize on Lexical Probability

	Estimate	SE	t	p	Sig.
Intercept	-4.48	0.04	-110.40	0.00	***
DSST[a]	0.07	0.02	2.78	0.01	**
ContextWindowSize	0.84	0.02	49.26	0.00	***
Group	-0.18	0.05	-3.38	0.00	***
ParentalSES[b]	-0.03	0.02	-1.67	0.10	.
SegmentLength	0.00	0.02	0.21	0.83	
ContextWindowSize*DSST	**0.01**	**0.01**	**0.90**	**0.37**	
ContextWindowSize*Group	**-0.06**	**0.02**	**-2.63**	**0.01**	*
ContextWindowSize*ParentalSES	0.00	0.01	-0.48	0.63	
ContextWindowSize*SegmentLength	-0.03	0.01	-3.91	0.00	***

Supplementary Table 18. Predictors of interest are shown in bold. [a]DSST = Digit-Symbol Substition Test score (mean of oral and written scores); [b]ParentalSES = Parental Socioeconomic Status.

Model Structure: LexicalProbability ~ ContextWindowSize*DSST + ContextWindowSize*Group + ContextWindowSize*ParentalSES + ContextWindowSize*SegmentLength + (1 + ContextWindowSize | Participant) + (1 + ContextWindowSize | Word)

Effects of TMT (Part 2) Score, Group, and ContextWindowSize on Lexical Probability

	Estimate	SE	t	p	Sig.
Intercept	-4.44	0.07	-64.11	0.00	***
TMT[a]	-0.04	0.05	-0.80	0.42	
ContextWindowSize	0.83	0.02	51.19	0.00	***
Group	-0.25	0.09	-2.70	0.01	**
ParentalSES[b]	-0.08	0.04	-1.82	0.07	.
SegmentLength	-0.04	0.03	-1.41	0.16	
ContextWindowSize*TMT	**-0.01**	**0.01**	**-1.16**	**0.25**	
ContextWindowSize*Group	**-0.06**	**0.02**	**-2.72**	**0.01**	**
ContextWindowSize*ParentalSES	-0.01	0.01	-1.04	0.30	
ContextWindowSize*SegmentLength	-0.03	0.01	-3.70	0.00	***

Supplementary Table 19. Predictors of interest are shown in bold. [a]TMT = Trail-Making Test, Part 2, score [b]ParentalSES = Parental Socioeconomic Status.

Model Structure: LexicalProbability ~ ContextWindowSize*TMT + ContextWindowSize*Group + ContextWindowSize*ParentalSES + ContextWindowSize*SegmentLength + (1 + ContextWindowSize | Participant) + (1 + ContextWindowSize | Word)

Effects of Category Fluency Score, Group, and ContextWindowSize on Lexical Probability

	Estimate	SE	t	p	Sig.
Intercept	-4.39	0.07	-61.04	0.00	***
CategoryFluency[a]	-0.04	0.05	-0.82	0.42	
ContextWindowSize	0.84	0.02	49.22	0.00	***
Group	-0.32	0.10	-3.27	0.00	**
ParentalSES[b]	-0.08	0.04	-1.85	0.07	.
SegmentLength	-0.04	0.03	-1.18	0.24	
ContextWindowSize*CategoryFluency	**-0.01**	**0.01**	**-0.72**	**0.47**	
ContextWindowSize*Group	**-0.07**	**0.02**	**-3.21**	**0.00**	**
ContextWindowSize*ParentalSES	-0.01	0.01	-1.22	0.23	
ContextWindowSize*SegmentLength	-0.03	0.01	-3.60	0.00	***

Supplementary Table 20. Predictors of interest are shown in bold. [a]Category Fluency score (number of exemplars); [b]ParentalSES = Parental Socioeconomic Status.

Model Structure: LexicalProbability ~ ContextWindowSize*CategoryFluency + ContextWindowSize*Group + ContextWindowSize*ParentalSES + ContextWindowSize*SegmentLength + (1 + ContextWindowSize | Participant) + (1 + ContextWindowSize | Word)

8.5 An exploration of potential non-linearity in the Disorganization*ContextWindowSize effect

There were a relatively small number of patients with higher thought disorder scores, and so it was unclear whether the interaction between Disorganization and ContextWindowSize was truly linear. To explore this relationship further, we performed two additional analyses: one in which we operationalize Disorganization as a binary variable based on a median split of patients' scores, and one in which we include the square of Disorganization as an additional predictor in the original analysis. We present these results below.

In the first of these two exploratory analyses, we show that the interaction between Disorganization and ContextWindowSize remains significant when we treat Disorganization as a

binary variable - that is, whatever the true shape of the relationship, we can be confident there *is* a linear component by which high-disorganization patients, in general, exhibit a lesser increase in lexical probability with increasing context, relative to low-disorganization patients.

Effect of Median-Split Disorganization and ContextWindowSize on Lexical Probability

	Estimate	SE	t	p	Sig.
Intercept	-4.67	0.05	-90.38	0.00	***
MedianDisorganization[a]	-0.24	0.10	-2.41	0.02	*
ContextWindowSize	**0.78**	**0.01**	**64.21**	**0.00**	***
ParentalSES[b]	-0.08	0.05	-1.61	0.11	
SegmentLength	-0.04	0.03	-1.21	0.23	
MedianDisorganization[a]*ContextWindowSize	**-0.06**	**0.02**	**-2.66**	**0.01**	*
ContextWindowSize*ParentalSES	0.01	0.01	0.70	0.49	
ContextWindowSize*SegmentLength	-0.03	0.01	-3.30	0.00	**

Supplementary Table 21. Predictors of interest are shown in bold. [a]Categorical: TLI score below median = -.5; TLI score above or equal to median = .5; [b]ParentalSES = Parental Socioeconomic Status.

Model Structure: LexicalProbability~ ContextWindowSize*MedianDisorganization + ContextWindowSize*ParentalSES + ContextWindowSize*SegmentLength + (1 + ContextWindowSize | Participant) + (1 + ContextWindowSize | Word)

In the second of these exploratory analyses, both the Disorganization*ContextWindowSize and Disorganization^2*ContextWindowSize interactions were significant, indicating that the relationship may have *both* linear and non-linear components. However, we interpret this post-hoc analysis cautiously and leave future studies with larger sample sizes to flesh out the true shape of this relationship.

Linear and Squared Effects of Disorganization and ContextWindowSize on Lexical Probability

	Estimate	SE	t	p	Sig.
Intercept	-4.69	0.05	-85.41	0.00	***
Disorganization	-0.20	0.08	-2.70	0.01	**
Disorganization^2	0.04	0.02	1.99	0.05	.
ContextWindowSize	**0.78**	**0.01**	**63.36**	**0.00**	***
ParentalSES[a]	-0.07	0.05	-1.34	0.19	
SegmentLength	-0.04	0.03	-1.18	0.24	
ContextWindowSize* Disorganization	**-0.06**	**0.02**	**-3.39**	**0.00**	**
ContextWindowSize* Disorganization^2	**0.01**	**0.00**	**2.17**	**0.04**	*
ContextWindowSize*ParentalSES	0.01	0.01	0.96	0.34	
ContextWindowSize*SegmentLength	-0.03	0.01	-3.16	0.00	**

Supplementary Table 22. Predictors of interest are shown in bold. [a]ParentalSES = Parental Socioeconomic Status.

Model Structure: LexicalProbability~ ContextWindowSize*Disorganization + ContextWindowSize*Disorganization^2 + ContextWindowSize*ParentalSES + ContextWindowSize*SegmentLength + (1 + ContextWindowSize | Participant) + (1 + ContextWindowSize | Word)

9. References

Arbuthnott, K., & Frank, J. (2000). Trail making test, part B as a measure of executive control: validation using a set-switching paradigm. *Journal of Clinical and Experimental Neuropsychology, 22*(4), 518-528. doi: 10.1076/1380-3395(200008)22:4;1-0;FT518

Becker, T. E. (2005). Potential Problems in the Statistical Control of Variables in Organizational Research: A Qualitative Analysis With Recommendations. *Organizational Research Methods, 8*(3), 274-289. doi: 10.1177/1094428105278021

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*: O'Reilly Media, Inc.

Black, S., Gao, L., Wang, P., Leahy, C., & Biderman, S. (2021). GPT-Neo: Large scale autoregressive language modeling with meshtensorflow: Zenodo.

Boyce, V., & Levy, R. (2023). A-maze of Natural Stories: Comprehension and surprisal in the Maze task. *Glossa Psycholinguistics, 2*(1). doi: 10.5070/g6011190

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language, 116*. doi: 10.1016/j.jml.2020.104174

Brothers, T., Morgan, E., Yacovone, A., & Kuperberg, G. R. (2023). Multiple predictions during language comprehension: Friends, foes, or indifferent companions? *Cognition, 241*, 105602. doi: 10.1016/j.cognition.2023.105602

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., . . . Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 1877-1901. doi: 10.5555/3495724.3495883

Chen, S. F., & Rosenfeld, R. (2000). A survey of smoothing techniques for ME models. *IEEE Transactions on Speech and Audio Processing, 8*(1), 37-50. doi: 10.1109/89.817452

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience, 8*(8), 1117-1121. doi: 10.1038/nn1504

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Douglas, R. J., Martin, K. A. C., & Whitteridge, D. (1989). A Canonical Microcircuit for Neocortex. *Neural Computation, 1*(4), 480-488. doi: 10.1162/neco.1989.1.4.480

Frisson, S., Harvey, D. R., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language, 95*, 200-214. doi: 10.1016/j.jml.2017.04.007

Galton, F. (1907). Vox Populi. *Nature, 75*, 450-451. doi: 10.1038/075450a0

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Giulianelli, M., & Fernandez, R. (2021). *Analysing Human Strategies of Information Transmission as a Function of Discourse Context.* Paper presented at the Proceedings of the 25th Conference on Computational Natural Language Learning (CoNLL).

Hale, J. (2001). *A probabilistic Earley parser as a psycholinguistic model.* Paper presented at the Proceedings of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL '01), Pittsburgh, PA.

Hao, Y., Mendelsohn, S., Sterneck, R., Martinez, R., & Frank, R. (2020). Probabilistic predictions of people perusing: Evaluating metrics of language model performance for psycholinguistic modeling. *arXiv preprint, arXiv:2009.03954*.

Kalnins, A. (2018). Multicollinearity: How common factors cause Type 1 errors in multivariate regression. *Strategic Management Journal, 39*(8), 2362-2385. doi: 10.1002/smj.2783

Kuperberg, G. R., Brothers, T., & Wlotko, E. (2020). A tale of two positivities and the N400: Distinct neural signatures are evoked by confirmed and violated predictions at different levels of representation. *Journal of Cognitive Neuroscience, 32*(1), 12-35. doi: 10.1162/jocn_a_01465

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience, 31*(1), 32-59. doi: 10.1080/23273798.2015.1102299

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*(5947), 161-163. doi: 10.1038/307161a0

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126-1177. doi: 10.1016/j.cognition.2007.05.006

Liddle, P. F., Ngan, E., Caissie, S. L., Anderson, C. M., Bates, A. T., Quested, D. J., . . . Weg, R. (2002). Thought and Language Index: an instrument for assessing thought and language in schizophrenia. *British Journal of Psychiatry, 181*(4), 326-330. doi: 10.1192/bjp.181.4.326

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology, 88*, 22-60. doi: 10.1016/j.cogpsych.2016.06.002

Luke, S. G., & Christianson, K. (2018). The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behav Res Methods, 50*(2), 826-833. doi: 10.3758/s13428-017-0908-4

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2023). Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, 1-71. doi: 10.1162/nol_a_00105

Mollica, F., Siegelman, M., Diachek, E., Piantadosi, S. T., Mineroff, Z., Futrell, R., . . . Fedorenko, E. (2020). Composition is the Core Driver of the Language-selective Network. *Neurobiol Lang (Camb), 1*(1), 104-134. doi: 10.1162/nol_a_00005

Murray, H. A. (1943). Thematic apperception test.

Neill, E., Gurvich, C., & Rossell, S. L. (2014). Category fluency in schizophrenia research: is it an executive or semantic measure? *Cognitive Neuropsychiatry, 19*(1), 81-95. doi: 10.1080/13546805.2013.807233

Palaniyappan, L., Al-Radaideh, A., Mougin, O., Gowland, P., & Liddle, P. F. (2013). Combined white matter imaging suggests myelination defects in visual processing regions in schizophrenia. *Neuropsychopharmacology, 38*(9), 1808-1815. doi: 10.1038/npp.2013.80

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog, 1*(8), 9.

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin and Review, 3*(4), 504-509. doi: 10.3758/BF03214555

Schmidt, A. F., & Finan, C. (2018). Linear regression and the normality assumption. *Journal of Clinical Epidemiology, 98*, 146-151. doi: 10.1016/j.jclinepi.2017.12.006

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences, 121*(10). doi: 10.1073/pnas.2307876121

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Sharpe, V. (2023). *Chapter 4: Patterns of lexical predictability in healthy adults' speech: Evidence from GPT-2 (Taking it with a grain of salt and pepper: Spared and impaired use of contextual information in the language production of people with schizophrenia).* (Doctoral dissertation), Tufts University, Medford, MA.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition, 128*(3), 302-319. doi: 10.1016/j.cognition.2013.02.013

Stroop, J. R. (1932). Is the judgment of the group better than that of the average member of the group? *Journal of Experimental Psychology, 15*(5), 550-562. doi: 10.1037/h0070482

Szewczyk, J. M., & Federmeier, K. D. (2022). Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language, 123*. doi: 10.1016/j.jml.2021.104311

Taylor, W. (1953). 'Cloze' procedure: A new tool for measuring readability. *Journalism Quarterly, 30*, 415-433.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., . . . Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv*. doi: 10.48550/arXiv.2302.13971

Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale, Revised*. New York: Psychological Corporation.